A VECTOR AUTOREGRESSIVE MODEL FOR INTERPOLATING MISSING METEOROLOGICAL DATA FOR USE IN BUILDING SIMULATION

Alisha A. Kasam, Benjamin D. Lee, and Christiaan J.J. Paredis
Model-Based Systems Engineering Center, Georgia Institute of Technology, Atlanta, GA

ABSTRACT

Building performance simulation is increasingly used to aid in decision making about the design, construction, retrofit, operation, and maintenance of new and existing buildings. Such simulations require a complete set of meteorological data sampled at regular intervals. A data file with even a single missing measurement value becomes useless for simulation. Unfortunately, it is extremely rare to find such a perfect body of data. Measurement errors and sensor failure are frequent occurrences in meteorological data collection and are among a host of reasons for missing measurement values. To overcome this problem, simulation users may rely on Typical Meteorological Years (TMYs) instead of actual historical data, or they may apply an existing interpolation method to fill the gaps in historical data. Historical data is often preferable, since TMYs fail to account for atypical weather conditions. Clearly, this could lead to poor decision making when the decision outcomes are strongly affected by the occurrence of atypical conditions.

This paper presents a novel application of Vector Autoregressive Gaussian Interpolation (VGI) as a method for statistical interpolation between discrete weather-data points. A vector autoregressive model is first calibrated using site-specific meteorological data, and then used to determine the most likely value for one or more missing data points. The method is validated for several cities in the USA, and results show a significant improvement in accuracy relative to other interpolation methods.

INTRODUCTION

In 2011, the residential sector accounted for 22% of end-use energy consumption in the United States (U.S. Energy Information Administration, 2012). In spite of several advances in building technologies, average household energy consumption has remained relatively stable. Economic and environmental concerns about the lack of improvement in energy efficiency have driven increasing interest in off-grid, zero-energy homes.

However, research has shown that a major obstacle to their implementation remains the risk arising from the inherent variability in the weather (Hu, 2009). There is significant uncertainty in both energy production and consumption in zero-energy homes. Due to the dynamic variation in the availability of natural resources such as solar radiation and wind, actual energy generation can be very difficult to predict. Similarly, the demand side of the energy balance is also strongly impacted by the surrounding weather conditions. Even typical variation in the surrounding weather conditions has been found to result in variations of 5% in actual energy consumption (Bhandari, et al., 2012). Considering these findings, weather uncertainty analysis and risk-conscious design have gained attention, especially in the building simulation community (de Wilde, et al., 2002).

Risk-conscious design demands a weather-data set that can provide not only the typical weather conditions for a specific location, but also the probability of random, risk-related weather conditions (Hu, 2009). Various "typical" data sets derived from observed meteorological data, most notably the Typical Meteorological Year (TMY) in the United States and the Test Reference Year (TRY) in the United Kingdom, have become the standard sources of meteorological data for building performance evaluation (Aguiar, et al., 1999). According to the user's manuals for the TMY2 and TMY3 standards, these profiles represent typical rather than extreme conditions and are not suited for worst-case scenario design (Marion and Urban, 1995, Wilcox, et al., 2008). Additionally, since these typical data sets are themselves based on historical data, data filling for gaps of up to 2 hours was performed using linear interpolation. Linear interpolation is generally satisfactory when the variable concerned varies slowly compared to the sampling interval, but even then, it can smoothen data excessively (Gorman, 2009).

Although no one typical data set or method has proven to be consistently better than the others, the study by Bhandari shows that a building's projected overall annual energy consumption varies up to 7%

depending on the provided weather data. For just heating and cooling loads, variation between data sets increases up to 40%. It can be logically concluded that different weather variables influence different aspects of building performance, and in fact, for an individual weather variable, the difference in projected consumption using different weather-data sets can be as high as 90% (Bhandari, et al., 2012). These simulation differences emphasize the importance of selecting an appropriate weather profile for building simulation. In this paper, we seek to address the deficiency in current interpolation methods by introducing a vector autoregressive model for the statistical interpolation of meteorological data.

The rest of the paper is organized as follows: First, we introduce a time-series approach to the analysis of meteorological data, which serves as the basis for this work. Next, the statistical interpolation model is presented, followed by validation examples representing several climate regions in the United States. Results of the statistical interpolation are then compared to those from linear interpolation, a current standard. Lastly, we discuss advantages of each approach, identify opportunities for future work, and offer concluding remarks.

BACKGROUND

Although actual historical data can be an attractive choice for risk-conscious design, its many gaps and flaws motivate an interpolation method which preserves the variation found in natural phenomena. In the case of errors or missing values, linear, spline, or nearest-neighbor interpolation are commonly used methods to complete a data set (Canale and Chapra, 2002). These relatively simple predictors are based only on the immediately surrounding observations, and their accuracies greatly decrease when used to predict more than one consecutive missing point. In the event of a measurement apparatus malfunction, for example, larger gaps in data may occur, and it is common to substitute the values from the same hours of the previous day or simply seasonal averages at those hours (Chatfield, 2004, Marion and Urban, 1995, Wilcox, et al., 2008). This paper presents an interpolation model which improves the accuracy of estimated values by considering climatological trends for a particular location.

Several commercial vendors offer proprietary interpolation and forecasting services which they claim are more accurate than the standard methods used by many researchers. However, the cost, delivery time, and inaccessibility to the methodology present considerable disadvantages to the user. In contrast, we present a model that can be freely used and understood by researchers in the building simulation community and other fields.

In order to define long-term meteorological trends, the chronological weather set is analyzed as a time series, defined as "a collection of observations made sequentially through time" (Chatfield, 2004). Since we assume non-deterministic behavior, a vector autoregressive model is used to predict the most likely value based on previous data, and statistical tools are used to characterize the uncertainty or variability. In addition, a time series analysis enables detection of cross-phenomena trends in weather data. Autoregression is a common single-variable approach to modeling a data trend. However, when the value of a given variable depends not only on past values of that same variable, but also partly on the values of at least one other time series, vector autoregression allows the modeling of multivariate trends. For example, solar radiation and relative humidity (RH) are both strongly related to dry bulb temperature (DBT) (Hassan, 2009). The strong relationship between weather variables indicates that incorporating multiple meteorological phenomena could provide a significant boost in the accuracy of the model's predictions. It is beyond the scope of this paper to describe time series analysis in detail, but for a more in-depth discussion, the reader is directed to (Chatfield, 2004).

Before it can be systematically analyzed, the data must first be "cleaned". Cleaning consists simply of detecting incomplete or corrupt parts of the data, and either inserting placeholders or removing errors as necessary.

It must then be determined whether the cleaned series exhibits a trend and/ or seasonality. In nearly all locations, seasonality in the data must be adjusted in order to analyze underlying meteorological trends. The most common method to eliminate seasonality is to use some form of a moving average, which smoothes the data over a specified window (Chatfield, 2004). Another option with regularly sampled observations is seasonal differencing, where each value is subtracted by the value from exactly one year prior, so that the time series consists only of the difference in observations from year to year. In the next section we describe the Rosenblatt transformation, a normalization technique which not only removes seasonality, but also generates a statistical characterization of the phenomena.

ESTIMATION METHOD

In this section, we introduce the Vector-Autoregressive Gaussian Interpolation (VGI) method. This method is based on the assumption that the weather at a given time is correlated to weather at similar times, with a noise term accounting for the residual variation. Characterizations of vector autoregressive processes typically assume that the residual term is Gaussian distributed. As shown in the next section, meteorological phenomena do not generally follow a Gaussian distribution, meaning that the residuals from the linear vector autoregressive model will not either. We therefore use a Rosenblatt transformation from the original domain into a normalized domain such that the

distribution of the transformed weather is Gaussian. The steps to determine the underlying meteorological trends and interpolate missing values in the data using the VGI are summarized in Figure 1, and more detailed pseudo-code is provided in Appendix 1.

Collect Site-Specific Meteorological Data

The first step in developing a site-specific interpolation model is to collect a sufficiently large sample of meteorological data. A weather database made publicly available by the National Climactic Data Center of the U.S. Department of Commerce is the source for all data used in the validation examples presented in the next section. Although calibration of the model is mathematically possible based on only a few years of data, it is strongly advised to use larger data sets in order to ensure that long-term trends and variability are captured. The data set used for model calibration in the validation examples included 40 years of hourly-sampled weather observations.

Since the model builds on hourly data for each phenomenon across all years in the data set, all observations from the 29th of February in leap years are neglected. In addition, missing measurements or unreliable measurements (as indicated by an error code) should be marked so that they are not included in the calculations. Finally, since measurements were only recorded every 3 hours during some years, it is necessary to insert empty placeholders between observations to match the hourly sampling rate of the rest of the data set.

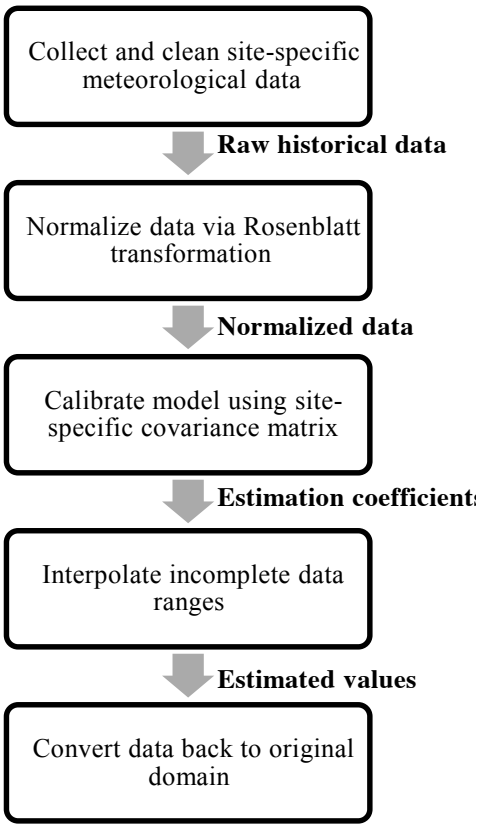

Figure 1. Steps in VGI Method

Normalize the Data

The VGI method assumes data to be realizations of a Gaussian process. However, it has been noted that historical meteorological phenomena do not generally follow a Gaussian distribution (Hong and Jiang, 1995, Lee, et al., 2012). Since such a characterization is the mathematical basis of the vector autoregressive model, the data is transformed from the original domain to the normalized domain using a Rosenblatt transformation (Rosenblatt, 1952). A Rosenblatt transform is generated and then applied for each hour of the time series such that at each hour, each phenomenon is characterized by a standard normal distribution.

In contrast to the procedures for removing seasonal variations referred to in the previous section, a Rosenblatt transform removes seasonality and also normalizes the distribution into a zero-mean unit-variance normal distribution. As visualized for a specific hour in Figure 2, the empirical CDF based on the historical data is first transformed into a continuous CDF using kernel-smoothening. It is then normalized by applying the standard-normal inverse CDF mapping. To provide more data to generate a smoother CDF, we chose to include the data from 24 and 48 hours prior to and following each hour. Although this may introduce a small bias, our validation experiments have shown that this improves the accuracy of the estimator.

The resulting range of normally approximated sample data [Y] and the statistical densities [F] computed at those values are stored for later use, when the normalized approximations are converted back to the original domain.

Calibrate Model for Location

Because the Rosenblatt transformation yields an approximately normally-distributed set of samples, we can assume that the missing values will also follow a normal distribution. The basis for the model then becomes a multivariate Gaussian distribution. Specifically, we assume that any data point in the normalized data set is statistically correlated to data points for all the measured phenomena at neighboring time instances. The joint probability distribution for this normalized data can then be represented as:

$$f(\boldsymbol{\phi}) = \frac{1}{(2\pi)^{\frac{nk}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{\phi}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu})\right) \quad (1)$$

where $\boldsymbol{\mu} \in \mathbb{R}^{nk}$ is a vector of expected values (which due to the application of the Rosenblatt transformation is near zero) for the normalized data $\boldsymbol{\phi}$, n is the number of meteorological phenomena considered, k is the number of time samples or lags from each phenomenon, and Σ is the corresponding covariance matrix. To determine the expression for the estimator of ϕ_1 given observations $\phi_2, \ldots, \phi_{nk}$,

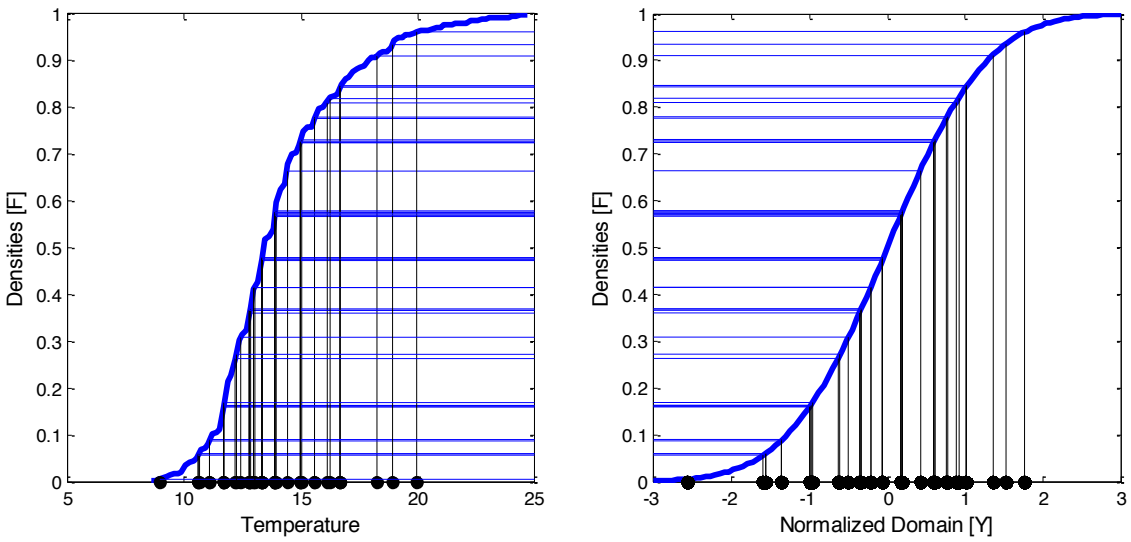

Figure 2. Rosenblatt Transformation for 4 PM on June 16, 1961-2011 in Seattle

the inverse covariance matrix is divided into sub-matrices:

$$\Sigma^{-1} = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \quad (2)$$

For the inverse covariance matrix constructed to solve for a single missing value, the size of A is $[1 \times 1]$, the size of B is $[1 \times (nk-1)]$, C is $[(nk-1) \times 1]$, and D is $[(nk-1) \times (nk-1)]$.

There are multiple methods for determining the covariance matrix from the data. One possibility is to determine a unique covariance matrix for each of the 8760 hours in the year. However, the number of available data points for a given hour is very limited: at most 40 data points for a 40-year data set. Validation experiments revealed that a data set of this size is insufficient to obtain good estimation results, especially for the case in which the dimensionality, nk, is large so that the covariance matrix may become rank deficient.

To overcome this lack of data, we instead determine a single covariance matrix in which all the 8760 hours of the year are combined. Since the data have been normalized and seasonality has been removed, all hours of the year can be considered jointly.

Estimate Missing Values

After determining the covariance matrix using a portion of the data for calibration, an expression can be derived to estimate the missing values in the remainder of the data set. Given the assumption that the normalized data can be represented adequately using the multivariate Gaussian model in Equation 1, an equation can be derived for determining the conditional probability distribution for a missing point ϕ_1, given the lags or neighboring values.

$$f(\phi_1|\phi_2,\dots,\phi_k,\dots,\phi_{n1},\dots,\phi_{nk}) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}}\exp\left(-\frac{(\phi_1-\mu_1')^2}{2\sigma^2}\right) \quad (3)$$

where μ_1' is the conditional expectation of $\phi_1$and σ^2 is the conditional variance of ϕ_1. Assuming zero mean to remove sample bias, (Equations 1 and 2) can be reduced to the statistical model (The proof derived by the authors is included in the Appendix; an alternate form of the proof can be found on page 116-119 of (Eaton, 1983).):

$$\mathrm{E}[\phi_1|\phi_2,\dots,\phi_k,\dots,\phi_{n1},\dots,\phi_{nk}] = -\frac{1}{A}\times C^T\begin{bmatrix}\phi_2\\ \vdots\\ \phi_k\\ \vdots\\ \phi_{n1}\\ \vdots\\ \phi_{nk}\end{bmatrix} \quad (4)$$

The estimation parameters or coefficients, $-\frac{1}{A}\times C^T$, characterize the relationship between k observations for all n phenomena. For example, in the validation experiments presented in this paper, we estimate temperature by finding estimation coefficients for both temperature and humidity lags.

An alternate version of the model includes a "greedy algorithm" to search for the set of lags within a specified range that can best predict the value at any given hour. The algorithm exhaustively searches through the range of possible lags to find the j predictors resulting in the lowest Mean Square Error, where j is also specified by the user. However, it can be seen in Figure 3 that the first one or two hours immediately preceding and following each data point provide the most correlation information in almost all cases, so the additional complexity required for the greedy algorithm lag search was not considered worthwhile. The first hour on each side exhibit a

strong correlation, with hours beyond the third exhibiting weak correlations which are likely influenced primarily by noise. While this closely resembles linear interpolation in the normalized domain, the inclusion of humidity data reveals the more complex relationships across phenomena, which will be discussed in more detail in the validation section.

Convert Data Back to Original Domain

The final step in the VGI algorithm is to convert the interpolated data from the normalized domain back to its original domain through the inverse Rosenblatt Transformation. The normalized data set is mapped through the standard normal CDF and then again through the kernel-smoothed inverse CDF generated from the stored values [*Y*] and [*F*].

RESULTS

This section introduces four examples to evaluate the model's performance given 51 years of historical weather data, sampled hourly at the following airports: Seattle (SEA), Albuquerque (ALB), Atlanta (ATL), and Minneapolis (MSP). Huang's study shows that weather variation has the greatest impact on energy consumption in heating-dominated locations, the least impact in cooling-dominated locations, and the most variable impact in balanced heating-cooling locations (Huang and Crawley, 1996). In order to evaluate the model's robustness, the sites at which the validation studies are performed have been selected to represent each of these three climatic categories. Seattle and Atlanta both require significant amounts of both heating and cooling, but are located at opposite corners of the country. Albuquerque is a cooling-dominated location, and Minneapolis is heating-dominated.

For each of these cities over the years 1961 to 2011, the model is used to estimate temperature by including humidity data and selecting only the immediately preceding and following hour as lags. To verify that over-fitting did not occur, the first 40 years in the series were reserved as calibration data, and each value in the remaining 11 years was estimated, one at a time, using the VGI method.

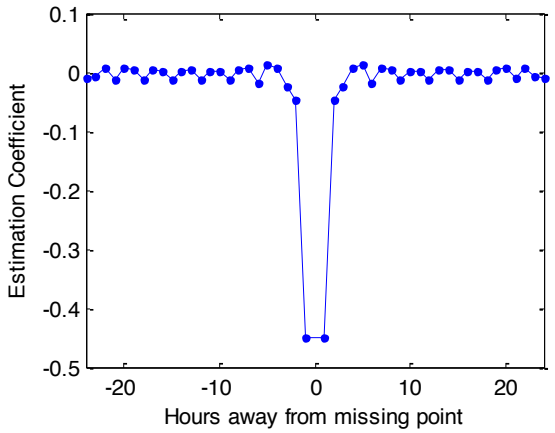

Figure 3. Estimation Coefficients over ± 1 day

Table 1 compares the temperature and humidity estimation coefficients for each of the cities. T_{-1}, T_{+1}, H_{-1}, and H_{+1} represent the coefficients for the temperature and humidity measurements taken at the hours immediately preceding and following the missing data point, and H_0 is the coefficient for the humidity value at the same hour as the temperature value being estimated.

As observed in the previous section, the contribution of the temperature lags is equivalent to linear interpolation in the normalized domain. In fact, when only temperature data from the hours immediately before and after are used for the VGI model, the estimated results are often similar to or even slightly less accurate than those obtained with linear interpolation. When humidity data is added to the covariance matrix, the average accuracy of the estimates improves dramatically, as shown in the next section. The significant differences in the humidity coefficients between cities may be explained by climate differences. However, it is worth noting that the weight of humidity data is approximately the same in Seattle and Atlanta, both considered balanced heating-cooling locations. In all four cases, the symmetry of the humidity coefficients results in a sum very close to zero. Furthermore, it appears that the coefficients correspond to an estimator for the second derivative of humidity with respect to time.

For each location, the humidity estimation coefficients for H_{-1}, H_0, H_{+1} are almost exactly factors of -1, 2, -1, corresponding to the second difference quotient approximation of the second derivative (Canale and Chapra, 2002). Figure 4 shows that for these coefficients, when humidity is either constant or varies linearly over a span of several hours, the contribution from humidity data is zero. Only when the data follows a curvature (e.g., near a peak or a valley) does the humidity data contribute to the temperature estimate. This holds true for any case where the humidity estimation coefficients sum to zero and display symmetry. Assuming humidity is well correlated with temperature, the curvature in the temperature can be estimated from the curvature of the humidity so that the humidity data improves the overall estimation quality.

Table 1.
Estimation Coefficients

	SEA	**ATL**	**ALB**	**MSP**
$\mathbf{T_{-1}}$	-0.500	-0.500	-0.498	-0.504
$\mathbf{T_{+1}}$	-0.498	-0.500	-0.498	-0.498
$\mathbf{H_{-1}}$	-0.218	-0.236	-0.327	-0.154
$\mathbf{H_0}$	0.421	0.458	0.655	0.286
$\mathbf{H_{+1}}$	-0.201	-0.221	-0.326	-0.130

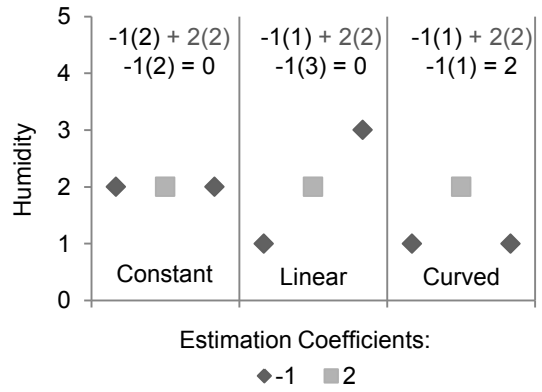

Figure 4. Second Derivative of Humidity

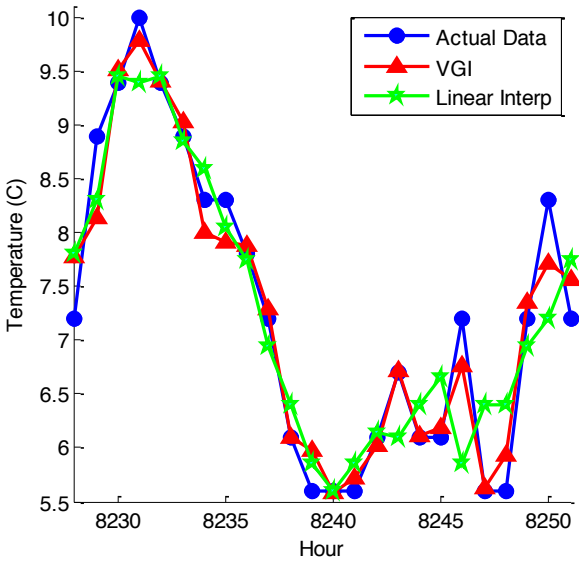

Figure 5. Prediction Comparison

VALIDATION

In addition to the experiments estimating temperature values in the four test cities, simple linear interpolation was performed on the same 11 validation years. In Figure 5, actual historical data spanning approximately one day in Seattle is plotted with the estimations from both interpolation methods. As indicated by the results from the second derivative of humidity, VGI performs significantly better at peaks and valleys than linear interpolation. However, this alone does not prove that the model performs better in general. Its overall estimation precision is emphasized in Figure 6, which compares the CDFs of the VGI error and the standard linear interpolation error, calculated for every estimated value in an 11-year period in Seattle. The long-term trend reveals that the model error shows significantly less variation than the linear interpolation error.

The mean of the residuals between the estimations and the known historical values was also compared between VGI and linear interpolation. The absolute value of each residual, e_{VGI} and e_{lin}, is calculated as shown in (Equation 5):

$$e_{\mathrm{VGI}} = |T - T_{\mathrm{VGI}}| \text{ and } e_{\mathrm{lin}} = |T - T_{\mathrm{lin}}| \qquad (5)$$

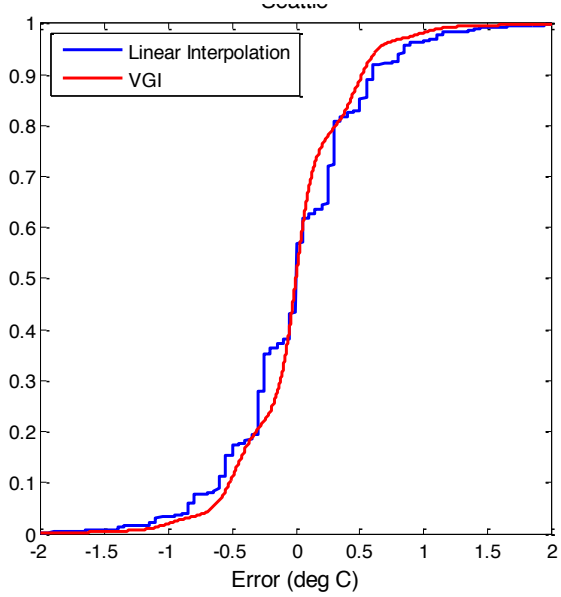

Figure 6. Error Comparison

Table 2.
Error Reduction of VGI Model

	$E[e_{\mathrm{VGI}}]$ (˚C)	$E[e_{\mathrm{lin}}]$ (˚C)	VGI Error Reduction
WA	0.3028	0.4481	48%
GA	0.3239	0.4408	36%
NM	0.4370	0.6166	41%
MN	0.3526	0.4146	18%

where T represents the actual temperature value, T_{VGI} is the temperature estimated using VGI, and T_{lin} is the temperature estimated using linear interpolation. The mean values of e_{VGI} and e_{lin} are calculated over all hours contained in the 11-year validation period. Table 2 displays the reduction in the mean residual using the VGI model relative to basic linear interpolation.

The prediction comparison, error comparison, and error reduction analysis together confirm that the VGI model outperforms linear interpolation over a range of climates.

CONCLUSION

The research shows that the hour directly before and after each point in time provides the most predictive information. What distinguishes the model from linear interpolation is its ability to determine estimation coefficients for other meteorological phenomena at the corresponding hours. Adding humidity data to the analysis provides a significant increase in the accuracy of temperature estimations for the cities validated thus far.

The multivariate interpolation framework presented in this paper can be applied to any geographical

location with long-term historical data. For the four validation examples studied thus far, the reduction in relative error for temperature estimations ranges from 18% to 48%.

Further development is needed to enable the model's use for predicting two or more consecutive missing values. A possible method handles each combination of consecutive missing values across all considered phenomena as a special case. For each case, the surrounding time samples which best predict the missing values are identified as the set of best lags. Once the lags are defined for all special cases such that they hold true for any location, the model can be applied iteratively over the data set, using the corresponding set of lags where each special case is identified. Additional further research must determine the extent to which VGI can estimate combinations of missing points. For instance, measurements were only recorded every 3 hours during the years 1971 and 1972 in Seattle. There may not be enough data available for the model to reliably estimate the missing values over this period.

Applications outside of building simulation and design provide further motivation for an improved interpolation tool for meteorological data. Reliability of automated weather stations has improved, but failures still occur, so techniques to fill gaps in data are still needed. Historical weather data has become important for prediction of insect- and disease-related crop damage, probability of forest fires, decision-making for irrigation, and numerous other agricultural applications (Acock and Pachepsky, 2000).

Experiments to evaluate the performance of the model have also demonstrated its usefulness in identifying outlier data points. When the model is applied over a range of data, it should produce a profile that closely resembles the empirical data, so that any large differences between the measured data and the model's estimate can serve to flag either a particularly irregular weather event or a flaw in the data itself. This ability to assist in the detection of historical sensor failure demonstrates the model's broader applications in other research domains as well.

NOMENCLATURE

$[Y]$ = normally approximated sample data
$[F]$ = densities at sample data
n = number of meteorological phenomena used
k = number of hourly lags
Σ = Covariance matrix
ϕ = Value of a meteorological phenomenon at a given time

APPENDIX

(Equation 3) gives the expected value of a missing measurement. It is derived from the correlated Gaussian model (Equation 1) as follows:

$$f(\boldsymbol{\phi}) = \frac{1}{(2\pi)^{\frac{nk}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{\phi}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu})\right)$$

$$f(\phi_1|\phi_2,\dots,\phi_k,\dots,\phi_{n1},\dots,\phi_{nk}) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}}\exp\left(-\frac{(\phi_1-\mu_1')^2}{2\sigma^2}\right)$$

where $= \begin{bmatrix} T_0 \\ T_{-1} \\ T_{+1} \\ H_{-1} \\ H_0 \\ H_{+1} \end{bmatrix} = \begin{bmatrix} \phi_1 \\ \widetilde{\boldsymbol{\phi}} \end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix} \mu_1' \\ \widetilde{\boldsymbol{\mu}} \end{bmatrix}$,

$$\Sigma^{-1} = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right];$$

μ_1' is the conditional expectation of ϕ_1 given $\widetilde{\boldsymbol{\phi}}$; σ^2 is the variance of ϕ_1; and because Σ^{-1} is symmetric, $B = C^T$.

Expanding:

$$(\boldsymbol{\phi}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu}) = \left[(\boldsymbol{\phi}-\mu_1)^T \quad (\widetilde{\boldsymbol{\phi}}-\widetilde{\boldsymbol{\mu}})^T\right] \times \begin{bmatrix} A & B \\ C & D \end{bmatrix} \times \begin{bmatrix} \phi_1-\mu_1 \\ \widetilde{\boldsymbol{\phi}}-\widetilde{\boldsymbol{\mu}} \end{bmatrix}$$

$$= (\phi_1-\mu_1)^T A(\phi_1-\mu_1) + 2(\widetilde{\boldsymbol{\phi}}-\widetilde{\boldsymbol{\mu}})^T C(\phi_1-\mu_1) + (\widetilde{\boldsymbol{\phi}}-\widetilde{\boldsymbol{\mu}})^T D(\widetilde{\boldsymbol{\phi}}-\widetilde{\boldsymbol{\mu}})$$

$$= \phi_1 A\phi_1 - 2\mu_1 A\phi_1 + \mu_1 A\mu_1 + 2(\widetilde{\boldsymbol{\phi}}^T C\phi_1 - \widetilde{\boldsymbol{\mu}}^T C\phi_1 - \widetilde{\boldsymbol{\phi}}^T C\mu_1 + \widetilde{\boldsymbol{\mu}}^T C\mu_1) + (\widetilde{\boldsymbol{\phi}}-\widetilde{\boldsymbol{\mu}})^T D(\widetilde{\boldsymbol{\phi}}-\widetilde{\boldsymbol{\mu}})$$

$$(\phi_1-\mu_1')^T\sigma^{-2}(\phi_1-\mu_1') = \phi_1\sigma^{-2}\phi_1 - 2\mu_1'\sigma^{-2}\phi_1 + \mu_1'\sigma^{-2}\mu_1'$$

Aligning similar terms:

(i) $\phi_1\sigma^{-2}\phi_1 = \phi_i^T A\phi_i \;\therefore \sigma^2 = \frac{1}{A}$

(ii) $-2\mu_1'\sigma^{-2}\phi_1 = -2\mu_1 A\phi_1 + 2\widetilde{\boldsymbol{\phi}}^T C\phi_1 - 2\widetilde{\boldsymbol{\mu}}^T C\phi_1$

(iii) $\mu_1'\sigma^{-2}\mu_1' = \mu_1 A\mu_1 - 2(\widetilde{\boldsymbol{\phi}}^T C\mu_1 + \widetilde{\boldsymbol{\mu}}^T C\mu_1) + (\widetilde{\boldsymbol{\phi}}-\widetilde{\boldsymbol{\mu}})^T D(\widetilde{\boldsymbol{\phi}}-\widetilde{\boldsymbol{\mu}})$

Assuming A is not rank deficient, solve (ii) for μ_1:

$$-\mu_1' A\phi_1 = -\mu_1 A\phi_1 + \widetilde{\boldsymbol{\phi}}^T C\phi_1 - \widetilde{\boldsymbol{\mu}}^T C\phi_1$$

$$\mu_1' = \mu_1 - (\widetilde{\boldsymbol{\phi}}-\widetilde{\boldsymbol{\mu}})CA^{-1}$$

This is generalized for any phenomenon $\boldsymbol{\phi}$ to solve for a single missing value ϕ_1:

$$\mathrm{E}[\phi_1|\phi_2,\dots,\phi_k,\dots,\phi_{n1},\dots,\phi_{nk}] = -\frac{1}{A}\times C^T\begin{bmatrix} \phi_2 \\ \vdots \\ \phi_k \\ \vdots \\ \phi_{n1} \\ \vdots \\ \phi_{nk} \end{bmatrix}$$

ACKNOWLEDGEMENT

This research is sponsored in part by the National Science Foundation under grant EFRI-SEED Award #1038248. This research is also partially sponsored by the Georgia Tech President's Undergraduate Research Award. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Georgia Institute of Technology.

REFERENCES

Acock, M., and Pachepsky, Y. A., 2000, "Estimating Missing Weather Data for Agricultural Simulations Using Group Method of Data Handling," *Journal of Applied meteorology*, **39**(7), pp. 1176-1184.

Aguiar, R., Camelo, S., and Gonçalves, H., 1999, "Assessing the Value of Typical Meteorological Years Built from Observed and from Synthetic Data for Building Thermal Simulation," *Proceedings of the 6th International IBPSA Conference ob Building Simulation'99 in Kyoto*, **2**, pp. 627-634.

Bhandari, M., Shrestha, S., and New, J., 2012, "Evaluation of Weather Data Sets for Building Energy Simulation," *Energy and Buildings*(0).

Canale, R. P., and Chapra, S. C., 2002, *Numerical Methods for Engineers*, Mc Graw Hill, New York.

Chatfield, C., 2004, *The Analysis of Time Series: An Introduction*, CRC press.

de Wilde, P., Augenbroe, G., and van der Voorden, M., 2002, "Design Analysis Integration: Supporting the Selection of Energy Saving Building Components," *Building and Environment*, **37**(8), pp. 807-816.

Eaton, M. L., 1983, *Multivariate Statistics: A Vector Space Approach*, John Wiley & Sons.

Gorman, R. M., 2009, "Intercomparison of Methods for the Temporal Interpolation of Synoptic Wind Fields," *Journal of Atmospheric and Oceanic Technology*, **26**(4), pp. 828-837.

Hassan, R., 2009, "A Comparison of the Accuracy of Building Energy Analysis in Bahrain Using Data from Different Weather Periods," *Renewable Energy*, **34**(3), pp. 869-875.

Hong, T., and Jiang, Y., 1995, "Stochastic Weather Model for Building Hvac Systems," *Building and Environment*, **30**(4), pp. 521-532.

Hu, H., 2009, *Risk-Conscious Design of Off-Grid Solar Energy Houses*, Thesis, College of Architecture, Georgia Institute of Technology, Atlanta.

Huang, Y. J., and Crawley, D. B., 1996, "Does It Matter Which Weather Data You Use in Energy Simulations?," *American Council for an Energy Efficient Summer Study*, Ernest Orlando Lawrence Berkeley National Laboratory, Pacific Grove.

Lee, B. D., Sun, Y., Hu, H., Augenbroe, G., and Paredis, C. J. J., 2012, "A Framework for Generating Stochastic Meteorological Years for Risk-Conscious Design of Buildings," in *SimBuild 2012*, Madison, WI.

Marion, W., and Urban, K., 1995, *User's Manual for Tmy2s: Typical Meteorological Years: Derived from the 1961-1990 National Solar Radiation Data Base*, National Renewable Energy Laboratory.

Rosenblatt, M., 1952, "Remarks on a Multivariate Transformation," *The Annals of Mathematical Statistics*, **23**(3), pp. 470-472.

U.S. Energy Information Administration, 2012, "Annual Energy Review 2011," vol. DOE/EIA-0384(2011).

Wilcox, S., Marion, W., and Laboratory, N. R. E., 2008, "Users Manual for Tmy3 Data Sets," National Renewable Energy Laboratory, Golden, Colorado, http://www.nrel.gov/docs/fy08osti/43156.pdf.